Goldstein, H. (2019). PISA and the globalisation of education: A critical commentary on papers published in AIE special issue 4/2019. *Assessment in Education*, *26*(19), 665-674. https://doi.org/10.1080/0969594X.2019.1674244

PISA and the globalisation of education:

A critical commentary on papers published in AIE special issue 4/2019

by

Harvey Goldstein, school of Education, University of Bristol

I congratulate the guest editors, Caro & Kyriakides, (2019), on an interesting collection of papers on PISA; its achievements and possibilities. Some of these articles will serve as useful reference works for future research. As has been the case since the start of PISA in 2000, its implementation and published results have often been controversial and some of this controversy is reflected in these papers. The editors themselves provide a brief summary of each paper and I will not attempt to replicate that. Rather I will focus on how far the papers increase our understanding of what PISA does and how much it contributes to scientific knowledge, and also the extent to which each paper's authors address some of the broad questions of validity and impact. I will start by offering critical comments on each paper and then provide a general review that will attempt to evaluate the global role that PISA continues to play.

LeRoy et al:

This paper is critical of the fact that information is generally unavailable about the extent of SEN provision and the performance of SEN students across countries. In part, it points out, this is because PISA does not allow countries to use 'test accommodation' procedures, such as allowing extra time for special needs students, to be used in reporting results. The paper does not go into any details of how SEN data might be analysed or what uses it might have if available, dwelling instead on current efforts by bodies such as UNICEF to produce internationally acceptable instruments, as well as asking OECD to make available what data they currently have.

The discussion in the paper is certainly useful, but it does not tackle the fundamental issue of what a universal definition of SEN might consist of. Since, to some extent at least, any definition of SEN is bound to have locally (country specific) relevant elements, this raises the question of whether a universal definition is either possible or desirable. What counts as special needs in one system, requiring for example particular institutional structures or special relationships with families of the students, may be quite different from requirements in other countries. It is another aspect of what might be described as the fundamental paradox of international comparative testing enterprises, namely that attempts to provide standardised definitions of test constructs or factors related to them, as with SEN, have the effect of encouraging disparate educational systems, with varying aims and objectives, to move towards a more homogenous grouping. As with public accountability systems more generally, what is assessed is influenced by how it is assessed. Whether this is what we want is something rarely debated in discussions of PISA and I will return to this issue later.

Spaull

The paper by Spaull looks at how the sampling procedures used by PISA can underestimate country performance. In particular, the fact that especially in poor countries, the 'out-of-school' population is often very large and in particular that changes over time in the relevant proportions, masks

legitimate comparisons. The analysis and discussion use data from Turkey from 2003 onwards. They author points out some technical errors in the simple adjustments claimed by OECD for out-of-school students and indicates that these underestimate the effect of high proportions out of school. The author points out that OECD reassurances about the small effects of being out of school should be treated with reservations. The author also points out that there has been little research analysis of this issue, citing the apparent outstanding performance of Vietnam which has been discussed without mentioning that Vietnam had the highest proportion of students not tested.

Spaull then describes a procedure that can be used to adjust for out-of-school potential bias. This essentially assumes that of those 15-16 year olds not in PISA none would have achieved level 2 on the maths and literacy scales – level 2 being defined roughly as functional literacy and numeracy. From separate survey data one can estimate the proportions not in schools and eligible for PISA. The author then simply multiplies the proportion in PISA not achieving level2 by the proportion not eligible. This, presumably establishes a *lower bound* for the true population proportion not achieving level 2 but it is very difficult to see that it approaches anything like an unbiased estimate for the proportion itself, since it is difficult to believe that all of those ineligible to be in PISA are not functionally literate. Certainly, the author provides neither empirical evidence, nor reasoned argument, why this should be so. Moreover, as the proportion of those eligible to be in PISA rose considerably over time from 2003, the relationship between such a lower bound and the true proportion would almost certainly change, ruling out any reasonable interpretation of changes in what the author refers to as a 'corrected estimate'. Spaull goes on to apply the same logic to the data split by socioeconomic group, with the same problems.

While the author certainly makes out a convincing case for retaining scepticism about interpretations of OECD data, and especially about changes over time, the proposed remedy hardly stands up. Nor does the suggestion that some form of 'imputation' might be used, since the required (strongly correlated) conditioning variables, notably prior achievement, are absent and it is very dubious as to whether any assumption about data being 'missing at random' could be sustained. In fact, this paper reveals one of the more serious weaknesses of the whole PISA enterprise, namely the absence of longitudinal data and I will return to that issue later.

Nagy et al.

This paper deals with the relatively technical, though important, issue of how the actual structure of the test instruments, especially item position, known as the position effect (PE), affects score responses.

A standard assumption underlying this paper is that individuals responding to a long test will tend to have a lower correct response to an item the nearer that item is located towards the end of the test. Explanations might be tiredness or, for example, lack of motivation. In other words, the probability of a correct response is a function of a presumed 'underlying ability', the presumed underlying characteristics of the item in question and the sequence position of the item in the test. Those students (just over 3%) who did not reach the end of the test were excluded – a somewhat strange thing to do since they would presumably be those students where the PE predicts very low probabilities towards the end of the test.

The PISA test design rotates booklets randomly among students within schools where a booklet will contain specific clusters of items for each domain. For any given domain, such as science, this means that each of the 7 science clusters appears in a different position (1 – 4) within a booklet and hence the position will vary across students. Thus, this is a study of the position effect on clusters of items

rather than individual items. School differences are incorporated by fitting random intercept multilevel models.

The authors present a structural equation model that assumes the mean response for a domain is an additive function of position, student 'proficiency' moderated by a set of covariates such as gender, SES and the school 'track' (low or Hauptschule, intermediate or Realschule, academic or Gymnasium and combined or Comprehensive). Each of these covariates was studied in turn for its effect on both the overall proficiency and the PE-proficiency relationship. Finally, an analysis was carried out to ascertain the effect that adjusting for PE would have on traditional analyses that ignored it. For this purpose the result for the PE of 1, where the cluster was in the first position, was assumed to represent the 'true' estimates. They conclude that Mathematics was least affected by PE and Reading most affected. They also found that school level effects differed in some cases from student level effects. They suggest that, although within a country PEs may not alter inferences substantially, they may do so between countries across which PE effects might vary. They also discuss more general interpretations of their results and the need to consider how PEs might affect inferences.

Despite one or two technical caveats; over the treatment of not-reached items and over the very small number of completed data sets analysed and combined using Rubin's rules (20 would be more appropriate for a multilevel model such as that being used), this is a carefully argued paper whose conclusions need to be taken seriously. One of the things it doesn't discuss in any depth is the possibility of redesigning PISA so that the burden on students is reduced, thereby presumably reducing the importance of PEs. This could, for example, entail increasing the number of booklets assigned in the overall matrix design, and an estimate of what would be required to obtain any given size of PE would be useful.

Robitzsch and Ludtke

For comparisons across time for different assessment test instruments, each needs to be transformed to a common scale. Thus, for two time occasions the standard procedure is to have a set of 'common' test items administered at each occasion that serve to 'anchor' each test and allow for example the second test to be rescaled so that the common items provide the same overall response. The assumption is that these common items can be interpreted as 'invariant' so 'aligning' the second (and subsequent) test with the first. The relationship between the non-common items and the common items is estimated, typically using an item response model, and this is what allows the alignment to be carried out. The analysis in this paper uses the estimated item parameters for each country and time and decomposes them into an item effect, a time effect and a country effect, each of which can then be estimated. It does not challenge the assumption of item invariance across time.

The linear models fitted then allow for estimation of uncertainty of the estimates attached to parameters measuring trends across time and country. The authors generate simulated data based upon PISA data in order to evaluate their proposed procedures and compare these with what are published by OECD. They also analyse actual data from PISA 2006 and 2009 surveys. Their general conclusion is that traditional PISA estimates of standard errors for trend estimates are too small. They also recommend that the number of common items is increased to reduce uncertainty.

Given its limitations, and the constraints imposed by the underlying anchoring assumption, this is a useful approach that PISA might wish to apply routinely in its reporting. What would be a very useful experiment for OECD to carry out would be one where different sets of common items were used for anchoring in order to shed light on the anchoring assumption itself. This could be achieved to

some extent by reanalysing existing data, but a more systematic selection of sets of common items could experiment with choosing items having different characteristics, possibly reflecting different dimensions of proficiency, etc. I shall have more to say about this later when discussing ways in which OECD could usefully broaden its scope and relax its assumptions about the rather simple item response models it currently operates.

Von Davier et al.

This paper also addresses the issue of measuring trends across time. The stated aim is to explore a more elaborate item response model than hitherto used by PISA to effect comparisons across countries and occasions. They restrict themselves, however, to the standard unidimensional model, justified it seems by the fact that 'the original (PISA) analysis also used a unidimensional model'!

The authors mention the difficulties associated with the 'item invariance' assumption due to interactions with curriculums etc and suggest that a limited number of such interactions can also be fitted, while retaining the main set of non-interactive terms that are used for scaling. The paper discusses the existing PISA procedures for equating items across countries and across time, based upon successive fitting and comparison of separate Rasch models within countries, across countries, within years and across years. Pointing out the weaknesses of this approach, especially in terms of its inability to deal with across-time biases arising from cumulative small changes, the authors propose a model that simultaneously fits data from different countries across different PISA testing occasions. One of the problems with this paper, along with the other papers in this collection, is that it does not set out the formal models used, so that in some cases it is not clear what assumptions are being made. In the present case, it appears that the model used constrains the parameters for each of the 'linking items', that is those that are assumed to be invariant across time, to be equal across time, and it also departs from the Rasch model by allowing for non-constant item 'discrimination' parameters.

Overall, they conclude that their new results, while being more stable, do align closely with existing PISA procedures. As one might expect the proposed joint model is also more easily able to study interactions between item, country and occasion, which is relevant to the interpretation of any trends. One conclusion is that some (small number of) items do not 'fit' well into the assumed item response model, and these can be given their own unique interaction parameter values, thus removing them from comparisons across time and country. Nevertheless, 'having your cake and eating it' in this fashion does remain problematic, although superior in general to the traditional PISA procedure of simply removing such 'dodgy' items. It does not, however, address the fundamental issue of comparability validity that I shall return to later.

Marchionni and Vazquez

These authors exploit the fact that around the fixed date when a child is allowed to enter school, children who share the same characteristics, except for their actual birth date which it is assumed to be independent of those characteristics, experience a year's difference in schooling. Thus, if we are interested in the pure causal effect of a year of schooling as compared with not going to school, we can simply compare these two groups of children, possibly adjusting for the actual small age difference involved. By 'causal effect' it is usually assumed that we mean the effect that would happen if, without changing anything else, a student stayed one year longer in school. It is as if the children had been assigned to the two groups at random. In such so called 'regression discontinuity (RD)designs' it is also typically assumed that there are no differential group (school or classroom)

effects, other than those that can be allowed for by including group fixed or random effects as in a multilevel model.

When the time around the fixed entry date is very small the design can be described as a 'sharp' RD design. As the authors point out, however, other factors such as repeating a year of schooling for low achievers can create interpretation difficulties. To incorporate such factors the authors adopt a 'fuzzy' RD design that incorporates a term for the probability of repeating etc. There are other issues that the authors attempt to deal with, such as the absence of an exact birth date and the form of the relationship assumed between age or birth date and outcome.

The authors present the results of fitting their model for 7 Latin American countries and show a strong effect of an extra year of schooling, in most cases for mathematics. The authors point out that PISA ignores different school entry cut-off dates and suggest that this can bias comparisons and they carry out an exercise to estimate what this would mean for their chosen countries. They also comment on the implications of policies that would extend the years of schooling in the light of their results. The authors contrast their results with more traditional methods that ignore the entry discontinuities.

There are, however, several issue with this analysis. The first concern is that, despite the hierarchical nature of the data, the analysis does not use multilevel models, so that significance levels will be overstated and confidence intervals too short. This is curious, since the authors do indeed mention the use of multilevel models when comparing with other analyses. Introducing a multilevel component into the model would also allow us to study, for example, whether the RD effect varied across schools. The second issue is that the existence of the discontinuity itself generates a confounding variable that is associated both with the treatment and with the outcome. This occurs because those students held back for a year because they are born after the entry date, will be the oldest students in their year group and those born just before the entry date will be the youngest in their year group. Since age-within-year-group is quite strongly associated with performance, this could at least partly explain the relationship with an extra year of schooling. In other words, simply extending the length of time that student spend in schools may not provide the improved performance indicated by the present analysis. It would be possible to include age-within-year-group as a variable in the analysis, but unless this is done, the results from the present paper need to be treated with caution.

Marksteiner et al.

Anchoring vignettes are an attempt to calibrate or 'anchor' differing individual interpretations of survey or other questions about themselves (or others), by constructing a 'common scale' that uses descriptions of the characteristics of hypothetically constructed individuals. These descriptions, since they are applied to a specific individual are assumed to be understood in the same way by the various survey respondents. The differences between the responses to the descriptions and the responses to questions about themselves are then assumed to provide a measure of differential item functioning (DIF) that then allows each individual's responses to be calibrated to a 'common scale'. Two principal assumptions are typically made in such an analysis, especially in the parametric modelling form that is widely used. The first is that each of the individuals will have the same underlying ranking for the vignette and the second that, apart from a simple underlying mean difference on an underlying 'latent' scale, individuals essentially have the same set of category assignments for the vignette and their own (subjective) individual ratings. The authors give as an example the concept of 'planning and organisation' where students were asked items on a Likert scale reflecting their own perceived levels of planning and organisation and then asked to rate the

same concept for three different hypothetical people who are described by how they plan and organise their work.

The data analysis consists of studying the structure of the responses in order to ascertain how far the basic assumptions behind the use of anchoring vignettes to adjust individual responses, can be sustained. In many ways the assumptions studied and the modelling procedures used parallel those used in item response models, although without the tendency of the latter to resist abandoning basic assumptions of comparability such as having a common set of 'equating' items for which strict comparability is always assumed.

The analyses carried out indicate that for two of the three anchoring vignettes used, there was little consistency in the way that students ordered the anchoring vignette constructs. They also report problems with the reliability of some of the scales (using Cronbach's alpha).

One of their principal conclusions is that anchoring did not help to establish 'measurement invariance' across countries. They comment that although PISA spends a great deal of time over translation and attempting to adapt to country specific features, language differences are still important and prevent international comparability of data. This backs up other work in this area (see for example, Blum et al., 2001). The authors appear to adopt a fairly pessimistic view of the use of anchoring vignettes and call for careful research before adoption. I shall return to the issue of comparability.

He et al.

While studying non-cognitive outcomes as Marksteiner et al, do, these authors utilise just the responses to self-report data to study comparability across countries.

To begin with this paper illustrates one of the problems with the technical language that is often used in describing the statistical techniques used to analyse test item or question scores. Thus, for example, in this paper the term 'bias' is used to indicate 'that scores from the assessment in different cultures reflect some cultural characteristics other than what the assessment is intended to measure'. To describe a lack of comparability in terms of 'bias' is pretty meaningless since no 'unbiased' value exists and the phrase 'intended to measure' leads to a circular definition. Further confusion is evident in the paper's distinction between 'Item response modelling' and factor analysis, whereas in fact they are conceptually the same, the only mathematical distinction being that in the former the response variable is an item – either binary or ordered which is assumed to reflect one or more underlying continuous characteristics, and in the second case the response is an actual measured continuous variable. In other words, the distinction is purely technical and formally independent of how results can be understood.

The starting point for the study reported in this paper is the existence in the 2015 PISA and the 2015 TIMSS studies of similar non-cognitive scales designed to elicit attitudes to the learning of science. The authors fit factor models to the 5 point Likert scales for each of the measurements, without, it seems, checking the statistical assumptions required such as multivariate normality. They carry out goodness of fit tests to attempt to establish whether it can be assumed that the scales are measuring the same constructs across countries and to establish whether the pattern of factor loadings is the same across countries. The usefulness of goodness of fit analyses is not discussed; these are generally a very poor way to study invariance since they typically have arbitrary thresholds for accepting a 'good' fit and often lack statistical power, whereas what is really needed are alternative hypotheses about the assumptions being made, most notably that the latent variables

(constructs) have a two or three dimensional rather than a 1-dimensional structure (see Goldstein and Browne, 2005).

The principal aim of the analyses is to see whether, using the correlations between the factor analyses carried out by the authors and science achievement, are closely related to the corresponding correlations for the Item response scaling scores used by PISA and science achievement. It isn't really clear why this is seen to be important, since the agreement is to be expected given that similar statistical models are used. In fact, the correlations themselves are only modest, and vary between countries. The paper makes little attempt to try and understand specific country differences and provides a minimal contribution to understandings of cross-country differences.

Some general observations

While the usefulness of many of these papers is questionable, they do all share certain, largely unexamined, assumptions. In this final section I want to try to extract some of the underlying issues about the general usefulness of International Large Scale Achievement (ILSA) programmes such as PISA and TIMSS, and the ways in which analyses of their data are reported.

Let me first discuss some technical issues (without being too technical about them) that repeatedly occur, with little sign of serious attempts to resolve them. To begin with the principal feature of these studies is that they are cross-sectional. This immediately limits the extent to which they can be used to make causal type statements. Yet, this is what most commentators and data analysts seek to do, whether by crudely linking country performance to educational policies or by attempting to exploit, with limited success, by-products of the sampling procedures as in the paper by Marchionne and Vasquez. Despite many commentators having drawn attention to this limitation, there has been hardly any movement towards providing longitudinal data, so that apart from anything else the scientific usefulness of such programmes remains highly problematic.

Secondly, the original procedure for providing individual level data for secondary analysis has hardly changed. This procedure, technically known as multiple imputation, but referred to by ILSA programmes using the term 'plausible values', is actually rather clumsy, and we know (see for example Goldstein et al., 2014) that they have certain severe modelling limitations such as being unable to provide unbiased model estimates when interactions are included. Furthermore, the scores that are released are ultimately derived from restricted item response models that seek to link items across rotated test booklets. There are alternative ways of using the 'raw' data derived directly from the booklets in a way that allows the data analyst flexibility in terms of combining components representing different curriculum subjects. With modern software packages this is relatively straightforward and, in particular, can properly incorporate multilevel data structures in ways that the existing data release is generally unable to do without introducing biases.

Thirdly, the existing ILSA programmes seem to be obsessed with the pursuit of 'comparability' across countries and across time. I have already commented on some of the caveats that should be made around all such attempts, see for example comments on the paper by LeRoy et al. I have elsewhere discussed more detailed technical limitations associated with country comparisons (Goldstein, 2004) and in particular why the simple one-dimensional comparisons utilised by PISA are often inadequate. It is important also to question why we should be so interested in this issue. To begin with, the term 'comparable' is vague with no formal definition other than that which may be implicit in terms of the assumptions made within the statistical models. To define 'comparability' as 'meaning the same thing' is tautological, yet as soon as we try to define it in terms of responses to test items or

questions, the whole issue of conditionality arises. If we have two countries where students of similar ages respond to the 'same' test we immediately have a problem with the term 'same' since aside from the obvious problem of translation, the familiarity of the test context or the extent to which what is being tested has been covered in any curriculum, arises. PISA evades the problem almost entirely by simply excluding ('dodgy') items that appear to exhibit unusual response patterns across countries, but this constitutes little more than a self-fulfilling prophecy. Likewise, trying to ensure comparability across time by keeping a set of 'calibration' or 'anchor' items that are reused and against which remaining items are scaled, typically involves making assumptions that ultimately rely upon a (subjective) judgement that a given set of anchor items 'mean the same thing' at two different occasions; an assumption that is not objectively testable.

It seems to me that ILSA programmes are supported by policymakers largely to provide comparisons so that they can claim 'success' for their own policies, if they happen to be in power, or to attack somebody else's policies, if they are not in power. Very few of the policy statements that quote results from studies such as PISA, whether made by national politicians or by PISA spokespeople, reflect any humility about the shortcomings of published data and the tentative nature of inferences that might be drawn.

Lastly, the very existence of PISA, and the obeisance afforded to it by its funders and supporting governments is moving the world towards a more homogeneous system of formal assessment and hence a homogeneity of the curriculums in different countries. The concepts ostensibly being tested are those considered important by a small number of technical and educational experts who are concentrated in a small number of countries. Diversity, it seems, is being sacrificed in favour or conformity where rankings and crude league tables dominate media coverage. The involvement of researchers carrying out secondary analyses, while it may be generally welcome, might also be questioned in terms of how much useful scientific knowledge accrues.

Thus, while I welcome many of these papers, on the whole I am disappointed. Many of the issues tackled, while of use in terms of showing how PISA can be improved, are relatively minor when set against the kinds of limitations I have described above. Many of the presentations themselves are also less than clear. For example, just three papers come anywhere near setting out clearly the formal statistical model being used (Robitzsch and Ludtke, Nagy et al., and Marchionne and Vazquez) while others are very difficult to understand because they omit details of the model they are using. There is also, of course, the question of whether journals such as AIE should be devoting quite so much space in special issues devoted to ILSA programmes when, in my view, it is now fairly clear that they have rather little to offer.

References additional to those cited in original articles.

Blum, A, Goldstein H & Guérin-Pace F. 2001. International Adult Literacy Survey (IALS): an analysis of international comparisons of adult literacy. *Assessment in Education,* Vol.8, No.2. pp225-246

Goldstein, H. (2004). International comparisons of student attainment: some issues arising from the PISA study, Assessment in Education: Principles, Policy and Practice, 11:3, 319-330, DOI: 10.1080/0969594042000304618

Goldstein H & . Browne W. 2005. Multilevel Factor Analysis Models for Continuous and Discrete Data. Contemporary Psychometrics, A. Maydeu-Olivares & J.J. McArdle (eds) Chap 14, pp 453-475.Lawrence Erlbaum Assoc. Publishers.

Goldstein, H., Carpenter, J. R. and Browne, W. J. (2014), Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. Journal of the Royal Statistical Society: Series A (Statistics in Society). 177(2), 553-564 doi: 10.1111/rssa.12022